

锋锐系列 DeepSeek推理微调一体机

CQ7258-A / 2U4卡PCIe



锋锐系列 DeepSeek 推理微调一体机型号为 CQ7258-A，是超擎数智专为 DeepSeek 70B 及以下模型开发的推理微调一体机。

锋锐系列 DeepSeek 推理微调一体机采用 AMD EPYC 9004 处理器，搭载 NVIDIA L20 GPU，2U4 卡 PCIe，采用“142”方案设计，搭载一颗 AMD EPYC 9004 处理器，连接 4 片 L20 GPU，专为满足企业 AI 基础设施的需求而打造，支持多达 4 个双槽主动或被动 GPU 可扩展配置。

DeepSeek 各参数模型推荐配置						
模型大小	1.5B	7B	8B	14B	32B	70B
模型精度	FP16	FP16	FP16	FP16	FP16	FP16
推荐台数	1	1	1	1	1	1

产品特性



强大性能

采用 AMD EPYC 9004 处理器提供高达 64 个核心和 128 个线程，适合处理需要大量并行处理能力的任务，如数据分析、科学模拟和复杂的计算工作负载。支持高速 DDR4 和新兴的 DDR5 内存技术，以及使用八通道内存配置，提供更大的内存带宽和容量，满足内存密集型应用的需求。



应用灵活

L20 GPU 配备第四代 Tensor Core 与第三代 RT Core，支持模型训练、推理、图形视觉等多种工作负载。



高性价比

采用 CPU 直通，无需 PCIe Switch 即可搭载 4 张 GPU，大大提升系统数据传输效率。

DeepSeek 性能测试

单台锋锐系列 DeepSeek 推理微调一体机性能输出				
模型大小	模型精度	output_max token	用户并发数量	token 吞吐量 (Token/s)
70B	FP16	512	80	870
32B	FP16	512	200	2100

产品规格

型号	CQ7258-A
处理器	1 个 AMD EPYC 9004 系列 (第四代 AMD EPYC 处理器, 最大热设计功率 400W)
内存	搭配 4 条 64G DDR5 内存条共 256G 内存, 可支持扩展至 12 条 DDR5 内存条共 3TB 内存, 内存类型: DDR5 4800/4400 RDIMM/RDIMM 3DS
存储	默认配置两块 480G SATA, 两块 3.84T NVME, 最大可支持 2 块 2.5 寸硬盘和 4 块 3.5 寸硬盘 (均支持 SATA、SAS 和 NVMe)
PCIE 扩展	后置: 4 个 PCIe x16 插槽, 用于双宽 GPU 卡 2 个 PCIe x16 插槽 1 个 PCIe x8 插槽 前置: 1 个 PCIe x4 插槽
网络	1 个板载千兆双口网卡 (电口) 和 1 个 RJ45 千兆管理端口
前置 I/O	4 个 USB3.2 端口
后置 I/O	2 个 USB3.2 端口 1 个 VGA 端口
尺寸	800mm*439.5mm*88.9mm
电源	1+1 冗余 2600W 80 PLUS 钛金 CRPS-R 电源

环境	工作温度: 10°C ~ 35°C 非工作温度: -40°C ~ 70°C 非工作湿度: 20% ~ 90% (非冷凝)
----	--

